BIOINFORMATICS ORIGINAL PAPER

Vol. 28 no. 13 2012, pages 1783–1789
doi:10.1093/bioinformatics/bts250

Databases and ontologies

Advance Access publication April 26, 2012

Semantic integration of physiology phenotypes with an application to the Cellular Phenotype Ontology

Robert Hoehndorf[1,*], Midori A. Harris[2], Heinrich Herre[3], Gabriella Rustici[4] and Georgios V. Gkoutos[1]

[1]Department of Genetics, University of Cambridge, Downing Street, Cambridge, Cambridge CB2 3EH, UK, [2]Department of Biochemistry; University of Cambridge, 80 Tennis Court Road, Cambridge CB2 1GA, UK, [3]Institute for Medical Informatics, Statistics and Epidemiology, University of Leipzig, Haertelstrasse 16-18, 04107 Leipzig, Germany and [4]European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, Cambridge CB10 1SD, UK

Associate Editor: Alex Bateman

ABSTRACT

Motivation: The systematic observation of phenotypes has become a crucial tool of functional genomics, and several large international projects are currently underway to identify and characterize the phenotypes that are associated with genotypes in several species. To integrate phenotype descriptions within and across species, phenotype ontologies have been developed. Applying ontologies to unify phenotype descriptions in the domain of physiology has been a particular challenge due to the high complexity of the underlying domain.

Results: In this study, we present the outline of a theory and its implementation for an ontology of physiology-related phenotypes. We provide a formal description of process attributes and relate them to the attributes of their temporal parts and participants. We apply our theory to create the Cellular Phenotype Ontology (CPO). The CPO is an ontology of morphological and physiological phenotypic characteristics of cells, cell components and cellular processes. Its prime application is to provide terms and uniform definition patterns for the annotation of cellular phenotypes. The CPO can be used for the annotation of observed abnormalities in domains, such as systems microscopy, in which cellular abnormalities are observed and for which no phenotype ontology has been created.

Availability and implementation: The CPO and the source code we generated to create the CPO are freely available on `http://cell-phenotype.googlecode.com`.

Contact: rh497@cam.ac.uk

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on January 20, 2012; revised on April 5, 2012; accepted on April 23, 2012

1 INTRODUCTION

Phenotype studies on all scales and levels of granularity are now an invaluable tool for functional genomics research. Phenotypes of targeted mutations in animal models are now systematically recorded to reveal the role of individual genes within a biological system. These phenotype studies now play a key role in translational research and are being used to reveal candidate genes for orphan diseases and to identify chemicals that may have effects on these diseases (Schofield *et al.*, 2011).

The large volume and diversity of phenotypes within different species and across multiple scales and levels of granularity necessitates the application of flexible strategies for managing and integrating data so that it becomes amenable to automated comparative analyses. To integrate biomedical data across heterogeneous information systems, biomedical ontologies are being developed (Smith *et al.*, 2007). An ontology is an explicit specification of a conceptualization of a domain and can be used to make the meaning of terms in a vocabulary explicit (Gruber, 1995; Guarino, 1998). They play a crucial role in the annotation of biomedical data and the integration of model organism databases (Bada *et al.*, 2004; Gene Ontology Consortium, 2010; Goble and Stevens, 2008).

Ontologies increasingly rely on the use of Semantic Web technologies (Berners-Lee *et al.*, 2001). The Semantic Web provides a stack of protocols and languages to include explicit semantics in websites. In particular, the Web Ontology Language (OWL) (Grau *et al.*, 2008) has been designed to express and share ontologies within the Semantic Web. OWL is a language based on description logics (a group of formal languages based on first-order predicate logic). Automated reasoners have been developed within the Semantic Web to perform complex operations on ontologies formulated in OWL. In particular, automated reasoners can verify an ontology's consistency and use deductive inference to perform powerful queries over ontologies. To benefit from automated reasoning and the rapidly increasing number of software tools that are being developed within the Semantic Web, most biomedical ontologies are now available in OWL or can be converted into an OWL-based representation (Hoehndorf *et al.*, 2010b; Horrocks, 2007).

In the domain of phenotypes, multiple ontologies have been developed. In particular, ontologies to characterize mammalian (Smith *et al.*, 2004), human (Robinson *et al.*, 2008), yeast (Engel *et al.*, 2010) and worm (Schindelman *et al.*, 2011) phenotypes are now available, while several more phenotype ontologies are under development. To benefit from automated reasoning, integrate phenotypes across species and reuse the content of anatomy and process ontologies, we have defined process classes using the framework of the Phenotypic Attribute and Trait ontology (PATO)

[*]To whom correspondence should be addressed.

(Gkoutos *et al.*, 2005). According to the PATO framework, a phenotype can be decomposed, using an Entity–Quality model, into an affected entity and a quality that characterizes *how* the entity is affected (Gkoutos *et al.*, 2005). Such decompositions have been created for several widely used phenotype ontologies (Gkoutos and Hoehndorf, 2011; Gkoutos *et al.*, 2009; Mungall *et al.*, 2010a), and are being applied together with methods for reusing knowledge contained in anatomy ontologies (Hoehndorf *et al.*, 2010a; Mungall *et al.*, 2010a).

Although the PATO framework is now successfully being applied to semantically integrate phenotypes across species, the diversity and complexity of phenotypes in which biological processes (BPs) and functions are impaired continues to limit the interoperability between phenotype ontologies. Major challenges for representing and integrating process phenotypes include establishing the link to the components of biological systems that have the capabilities to exhibit such a behaviour, and that attributes of processes are often measured *indirectly* and inferred from other attributes.

Here, we present the foundations for an ontology of process phenotypes. We present a theory in which several kinds of process attributes can be distinguished so that normal and abnormal physiology of biological systems can be formally characterized. We apply this theory to cellular processes and create the Cell Phenotype Ontology (CPO). The CPO is linked to reference ontologies for qualities, BPs, functions and cell components, and its prime application is the unification of phenotypes on the cellular level across different species as well as for annotation of cellular phenotypes in domains in which no such ontology exists.

2 SYSTEM AND METHODS

2.1 Gene ontology

The Gene Ontology provides a set of ontologies for molecular and cellular biology, originally designed to support structured annotations for genes and gene products in all species with respect to molecular function (MF), BP and cellular component (CC). MF and BP both describe processes, but at different spatiotemporal scales; in particular, BP includes processes that unfold within cells and within tissues and organs of multicellular organisms. Gene product annotations can be interpreted as identifying participants in the processes.

Over time, GO development has increasingly emphasized a normalized approach that includes supplementing existing human-readable text description with formally specified explicit definitions for GO classes. The formalization of GO is readily apparent in its representation of biological regulation.

Regulatory processes may regulate other processes, at either the MF or BP scale or biological qualities. GO accordingly includes three broad categories of regulation terms, regulation of MF, regulation of BP, and regulation of biological quality. The first two are explicitly defined entirely with respect to other GO terms, whereas the third comprises classes in which the regulated qualities are specified by terms from PATO (see below) or anatomy ontologies.

All GO regulation terms use one of three relations, **regulates**, **negatively_regulates** and **positively_regulates**, to link regulation terms to process or quality terms. The **regulates** relations are defined in terms of qualities: a regulatory process causes a change in magnitude to some quality, which in turn has an effect on the frequency, rate or duration of some other type of process. Effects that results in increases and decreases use **positively_regulates** and **negatively_regulates**, respectively (Mungall *et al.*, 2011). The existing ontology structure would also support the addition of subclasses to distinguish, for example, regulation of the rate of a process from regulation of its duration or time of onset.

2.2 PATO and the EQ model

PATO was envisaged and designed to provide a platform for allowing the integration of quantitative and qualitative phenotype-related information across different domains, levels of granularity and species (Gkoutos *et al.*, 2005). PATO is an ontology of phenotype qualities that form basic entities that we can perceive and/or measure such as colours, sizes, rates etc. One of its classification axes is based on the basic type of entity to which a qualities belongs, and PATO distinguishes between qualities of physical objects and qualities of processes.

PATO allows for the description of affected entities by combining various ontologies that describe the entities that have been affected, such as the various anatomical ontologies, GO (Ashburner *et al.*, 2000), the Cell Type Ontology (Bard *et al.*, 2005), SO (Eilbeck *et al.*, 2005), etc with the various qualities it provides for defining how these entities were affected. PATO can be used for annotation either directly in a so called post-composed (post-coordinated) manner or for providing formal (logical) definitions (equivalence axioms) to ontologies containing a set of precomposed (precoordinated) phenotype terms. For instance, to describe the decrease in the length of the sexual cycle of female animals, we can combine the PATO term *Decreased duration* (`PATO:0000499`) with the Gene Ontology term *Estrous cycle* (`GO:0042698`), while if such a term existed in a precomposed ontology (for example, the `MP:0009007` term from the Mammalian Phenotype) it could be used to provide an equivalence statement between that class and the above PATO-based description.

2.3 Axioms for physiology phenotypes

We implement our theory of physiology phenotypes using OWL, a formal language based on description logics. Using OWL, we formulate axioms that can be used by automated reasoners to infer additional information. Automated reasoning and the axioms we provide are intended to satisfy three aims. First, we use the axioms to infer information from background knowledge. In particular, we aim to automatically generate a class hierarchy of physiology phenotypes when an ontology of physiological processes, such as the GO, is provided as background knowledge. Our second aim is to provide interoperability with phenotype ontologies of other domains, including species-specific phenotype ontologies that have been formalized using the EQ method (Gkoutos *et al.*, 2005; Mungall *et al.*, 2010a). Finally, our third aim is to query physiology phenotypes based on physiological processes that are affected or based on the way in which they are affected.

Our three aims rely on the possibility of using automated reasoning over a resulting ontology of physiology phenotypes. However, OWL is an expressive formal language, and automated reasoning in OWL has a high computational complexity (reasoning in OWL 2.0 is 2NEXPTIME-complete). Consequently, due to the exponential increase in worst-case time complexity for automated reasoning in OWL, we would not be guaranteed to achieve our aims once we consider more than very few phenotype classes. In particular, using an ontology of the size of GO, with more than 22 000 classes for processes, as a foundation for constructing an ontology of physiology phenotypes would not allow us to achieve our aims through automated reasoning.

The OWL EL profile is a subset of OWL that significantly decreases the expressivity of OWL and the resulting time complexity of automated reasoning (Motik *et al.*, 2009). Highly efficient automated reasoners have been developed for OWL EL, which are capable of processing very large ontologies (Kazakov *et al.*, 2011). To achieve our aims and use automated reasoning for large ontologies, we limit ourselves to the OWL EL profile. As a consequence of the restriction to OWL EL, we cannot make use of negation (`not`), disjunction (`or`), universal quantification (`forall`) and several other types of operations in our axioms (Motik *et al.*, 2009).

The lack of expressivity in OWL EL requires a formulation of axioms so that the inferences we desire (i.e. the subclass relations resulting in the ontology's taxonomy) are maintained without using features of OWL that go beyond OWL EL expressivity. Consequently, we formulate *abnormality* and *absence* similarly to current formalizations of EQ-based phenotype ontologies, without the use of negation, disjunction or universal quantification (Mungall *et al.*, 2010a). A detailed description of the axioms we implement is available as Supplementary Material.

3 RESULTS

3.1 Attributes of processes

We develop a model of process attributes that is applicable for representations of physiology and related phenotypes. In principle, we distinguish between three different kinds of process attributes: the first are process attributes that arise directly from processes and include *Duration* and *Temporal location*; the second are attributes that arise from processes and their temporal parts and include *Frequency* and *Onset*; and the third are attributes that arise from processes and qualities of their participants, and include *Flow rate*s.

Attributes that can be directly linked to a process arise from processes' temporal extension. For example, a duration is an attribute that characterizes the temporal extent of a process and is similar to *Length*, *Area* and *Volume* for one-, two- and three-dimensional physical objects. A *Temporal location* positions the time interval at which the process occurs with respect to a reference coordinate system.

However, the majority of attributes that characterize processes are not based on these types of process attributes alone, but rather relate attributes of process participants with the duration of a process. In particular, a *Rate* typically refers to an attribute of some entity *with respect to an attribute of another entity*, and in the context of processes, rates often refer to attributes of a process participant with respect to the duration of the process. For example, a *Mass flow rate* refers to the *Mass* of a process participant with respect to the *Duration* of the process, that is, how much matter is moved (from one point to another) through the process. As a more complex example, a *Rate of change of position* refers to the *Distance* that an object is moved with respect to the *Duration* of the process.

However, not all rates of a process depend on attributes of the process participants. In particular, a *Frequency of occurrence* or *Event rate* refers to the number of occurrences of a type of process during a reference process. For example, a *Rate of heart beating* refers to the number of *Heart beating* processes that occur within a reference process (e.g. a process in which the heart participates with a duration of 1 min). Further attributes that depend on types of processes with regard to a reference process are *distribution patterns*, that is, how the occurrences of processes of a particular type are distributed within a reference process. For example, the heart may beat rhythmically or arrhythmically within a period of time (see Fig. 1).

Related to distribution patterns are *changing qualities* of processes. For example, the rate of heart beating may change (*increase* or *decrease*) throughout the course of a reference process. A simple analysis of *increasing* (*decreasing*) rates would be that the rate of a heart beating within the first half of a process is *lower* (*higher*) than in the second half of the process. To make such an assertion, we divide a process into two temporal parts. Mathematically, this process of subdivision can be iterated until processes occur within infinitesimally small time intervals.

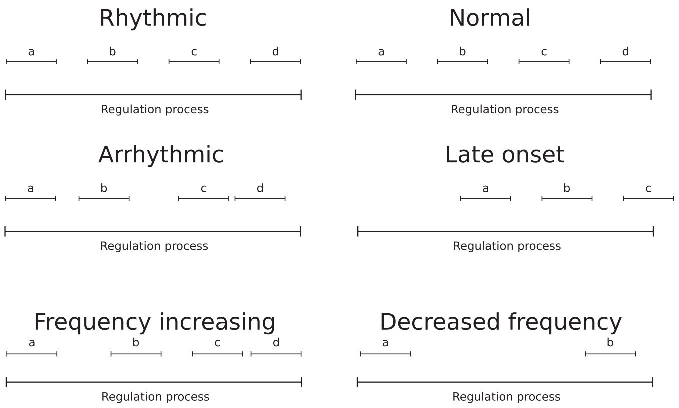

Fig. 1. Six examples of processes with non-comparative and comparative process attributes. We assume that the processes labelled a, b, c and d are all instances of the class of processes P. On the left side, three regulation (of P) processes are illustrated which exhibit non-comparative attributes. The first process has an attribute of *Rhythmic occurrence of P* because the instances of P are temporally equidistantly distributed. The second example shows an *Arrhythmic occurrence of P*, and the third examples shows an *Increasing frequency of P*. A regulation process with an *Increasing frequency of P* attribute is a process in which the value of the *Frequency of occurrences of P* attribute is lower in the first half of the process than in the second half. The right side of the figure illustrates comparative phenotypic descriptions of processes. On the upper right, the *normal* reference process is shown. The second example illustrates a *Late onset of P*, i.e. the attribute that P processes begin later than *normal* processes. Finally, the lower right illustrates a *Decreased frequency of P*, since fewer processes of the type P occur within the reference process than *normal*.

Although some processes can be subdivided indefinitely while retaining certain kinds of attributes, others cannot. Examples of processes that can be divided include *Continuous movement* or *Mass flow* processes, for which all parts have a *Speed* or *Flow rate* attribute. On the other hand, some processes can be subdivided into stages of activity and stages of inactivity (with respect to a particular process type) and cannot arbitrarily be divided. For example, a process of *Heart beating* has periods of activity in which a heart beat occurs and periods in which no heart beat process occurs. Consequently, not all parts of the process have a *Heart rate* (*Rate of heart beating*) attribute.

We may further attribute a *Frequency* or *Rate* to an object instead of a process. For example, a heart that beats *now* with a frequency of 80 bpm, or a car that is moving at a speed of 180 km *at a particular point in time* (e.g. as observed with a speed camera) can be considered attributes of the objects (the heart or the car), not attributes of the processes in which the objects participate. However, these are *different* kinds of attributes. Rates, when considered as attributes of objects, may be explicitly defined using rates of processes. For example, the heart beating frequency of a particular heart h at a time point t is the frequency of a reference heart beating process in which h participates. Such a reference process is necessary to obtain a value for a frequency even when no *Heart beating* process is occurring. However, the frequency is only an attribute of the heart in virtue of such a reference process in which *Heart beating* is actually occurring. This reference process can be uniquely determined for processes such as *Continuous movement*, where the rate of an object at a time t is the rate of the infinitesimally small process that occurs around t. The reference process is ambiguous for processes such as *Heart beating*, and the reference process must be explicitly stated.

3.2 Cell phenotype ontology

Although our considerations about process attributes are only the beginnings of a full-fledged theory, we have derived several phenotype formalization patterns and a high-level taxonomic structure of process-based phenotypes. To evaluate our approach, we created the CPO by automatically applying our patterns to the GO.

Phenotypes in the CPO are either based on structural abnormalities or abnormal physiology involving cells or cell components. Structural abnormalities in the CPO are based on GO-CC hierarchy. GO-CC contains 2918 classes for cell parts (including *Cell*) and extracellular components of cells. For each CC class *C* in the GO-CC, we create a new class labelled *C phenotype* in the CPO. For example, for the class *Mitochondrion* (`GO:0005739`) in the GO-CC, we create the class *Mitochondrion phenotype*.

Among the structural phenotype classes, we first distinguish between *normal* and *abnormal* phenotypes. An *Abnormal phenotype of C* is a phenotype of an organism that does not have a normal *C* as part, while a *Normal phenotype of C* represents the state in which an organism has a *normal C* as part.

Among the abnormal phenotypes that we include for all cell components listed in GO-CC, we distinguish *Abnormal morphology* and *Abnormal physiology* phenotypes. An *Abnormal morphology of C* is either the (abnormal) absence of required parts of *C*, the (abnormal) presence of additional parts, or abnormal qualities of *C* (Hoehndorf *et al.*, 2010a). For example, an *Absence of caveolae* (`MP:0004150`) would be a subclass of *Abnormal morphology of plasma membrane* in virtue of caveolae necessarily being part of the *Plasma membrane* (`GO:0005886`).

Abnormal physiology of a cell component refers to abnormal *functionality* of a cell component. We assume that a functionality of a cell component is (the potential for) a process in which the cell component is (causally) involved. We use the definitions of GO classes that were created based on lexical decompositions of GO class labels (Bada and Hunter, 2007; Mungall *et al.*, 2011; Ogren *et al.*, 2004) to identify the processes in which cell components are involved. For example, the definition of the GO class *Mitochondrial fission* (`GO:0000266`) is explicitly defined as an *Organelle fission* (`GO:0048285`) that **results-in-the-division-of** a *Mitochondrion* (`GO:0005739`). Based on this definition, we make the assumption that *Mitochondrial fission* is one of the functions of a *Mitochondrion* and that an *Abnormality of mitochondrial fission* is a subclass of an *Abnormality of mitochondrion physiology*.

Among abnormal physiology, we distinguish between abnormalities in a *single occurrence* of a cell component's functioning and an abnormal *pattern of multiple occurrences* of a cell component's functioning. For example, abnormalities in cell division resulting in *Aneuploidy* refer to abnormalities of *Cell division* processes, while an *Increased rate of cell division* refers to an abnormality in the pattern of occurrence of multiple cell division processes. In the CPO, we follow the GO and represent abnormalities in the pattern of occurrence of *X* as abnormalities of *Regulation of X* processes. In particular, an *Increased rate of cell division* is not an attribute of *Cell division* processes, but rather arises from the collection of all *Cell division* processes that occur within an organism at a given time. In the GO, *Regulation of X* processes refer to those processes that determine how often and in which way one or more *X* processes occur. Therefore, we assign the attribute of *Increased rate of cell division* to *Regulation of cell division* processes.

Single occurrences of processes can be abnormal in multiple ways, depending on the type of process. First, common to all processes is the quality of *Duration* and consequently, each process can have an *abnormal* (increased or decreased) duration. For example, a part of an organism may participate in an *Inflammatory response* (`GO:0006954`) that lasts longer than normal, that is, the organism has an *Increased duration of inflammatory response* phenotype. We define such a phenotype as a phenotype of an organism which has a part that participates in *Inflammatory response*, and this *Inflammatory response* process has an *Increased duration* (`PATO:0000498`).

The second common type of abnormality are abnormalities based on process participants in relation to the duration of the process. These include all kinds of *rates* such as *Mass flow rate*, *Energy flow rate* and *Velocity* (the rate of change of position). In each of these cases, an object participates in a process and a quality (or change of quality) of that object throughout the duration of the process is considered to form a new quality. If the process has participants that are distinguished into *inputs* and *outputs*, then a recurring pattern is that the amount of inputs or outputs that participate in the process can be *increased* or *decreased*. For example, an *Increased rate of cytoplasmic streaming* can be defined as an increased amount of inputs or an increased amount of outputs of a *Cytoplasmic streaming* process.

Finally, some objects may be divided into stages during which particular states of affairs obtain, and a process may be abnormal in that these states of affairs do not obtain at a particular stage. Notably, at the beginning and the end of a process, pre- and post-conditions may obtain that are abnormally changed in a process. For example, *Aneuploidy*—an abnormality during cell division at which the chromosomes do not separate properly between the two cells — may be considered the result of such an abnormality.

We implement the first two types of abnormality in the CPO. First, as a subclass of each *Abnormality of P* class, we create *Abnormal duration of P*, which in turn has *Increased duration of P* and *Decreased duration of P* as subclasses. Second, if we are able to identify *inputs* $I(P)$ or *outputs* $O(P)$ of the process P in the formal definitions of the GO, we automatically generate *Abnormality of* $I(P)$ *in* P as well as *Abnormality of* $O(P)$ *in* P. The left side of Figure 2 illustrates the schema of classes we generate for single process abnormalities.

The second type of abnormality in the CPO relate to abnormalities of *multiple* occurrences of some process X. According to the GO, *regulation* processes are processes that maintain or modify the occurrence of processes of a particular type. Following this convention, we call an abnormality of multiple occurrences of X *Abnormality of the regulation of* X.

A first kind of abnormality of regulatory processes are *abnormal temporal distribution patterns* of a process. In these abnormalities, the *way* in which processes of a particular kind are temporally distributed is abnormal. The most common abnormal distribution pattern is an increased or decreased frequency, and we use PATO's *frequency* class to define *Abnormal frequency of occurrence*

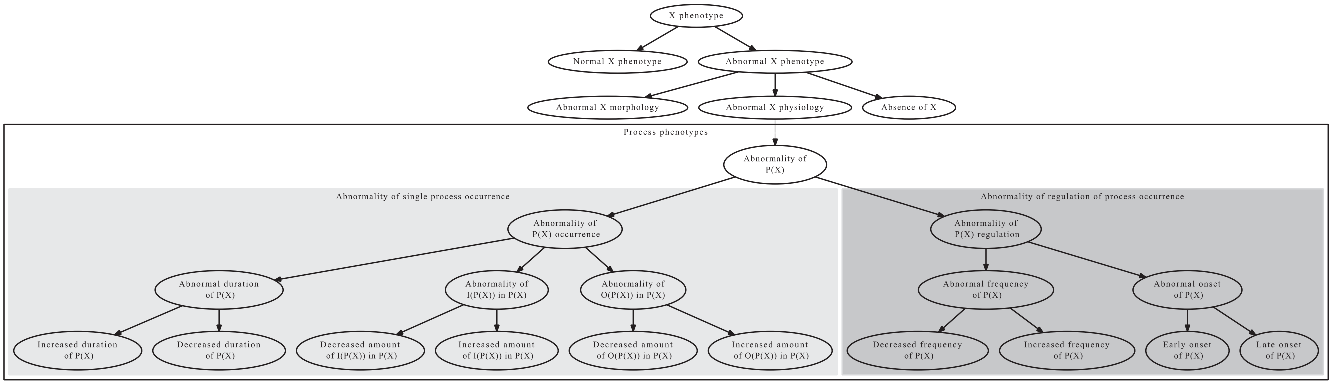
X phenotype
Normal X phenotype
Abnormal X phenotype
Abnormal X morphology
Abnormal X physiology
Absence of X
Process phenotypes
Abnormality of P(X)
Abnormality of single process occurrence
Abnormality of P(X) occurrence
Abnormal duration of P(X)
Abnormality of I(P(X)) in P(X)
Abnormality of O(P(X)) in P(X)
Increased duration of P(X)
Decreased duration of P(X)
Decreased amount of I(P(X)) in P(X)
Increased amount of I(P(X)) in P(X)
Decreased amount of O(P(X)) in P(X)
Increased amount of O(P(X)) in P(X)
Abnormality of regulation of process occurrence
Abnormality of P(X) regulation
Abnormal frequency of P(X)
Abnormal onset of P(X)
Decreased frequency of P(X)
Increased frequency of P(X)
Early onset of P(X)
Late onset of P(X)

Fig. 2. Overview over the taxonomic structure of CPO. The structure is based on a cellular component class X and the cellular processes $P(X)$ in which X is involved.

of X. For example, an *Abnormal frequency of occurrence of apoptosis* is defined as an abnormality of *Regulation of apoptosis* (`GO:0042981`) with respect to the *Frequency* (`PATO:0000044`) of *Apoptosis* (`GO:0006915`) occurrences.

There are further types of deviation from a distribution pattern. For example, a kind of process that is normally *rhythmic* can be abnormal in that it is *arrhythmic*. A typical example of this kind of process is *Heart beating* (`GO:0060047`), in which *Cardiac muscle contraction* (`GO:0060048`) processes occur in a rhythmic pattern. In *Cardiac dysrhythmia*, however, *Cardiac muscle contraction* processes occur arrhythmically, and we consider this to be an abnormality of the regulation of *Cardiac muscle contraction*. Although these abnormalities are often highly informative in clinical diagnostics and biological investigations, we usually lack the necessary information that is required to automatically determine meaningful types of abnormal distribution patterns.

A second kind of regulatory abnormalities is related to the *onset* of a process. With respect to a reference process, a particular kind of process may be *Delayed* (`PATO:0000502`) or *Premature* (`PATO:0000694`). For example, *Delayed apoptosis* refers to an abnormality of the *Regulation of apoptosis* in which apoptosis is induced later than normal. We use the PATO quality *Onset* (`PATO:0002325`) and its children *Delayed* and *Premature* to define these types of regulatory abnormality. Similarly, we use PATO's *Offset* (`PATO:0002324`) quality and its children to characterize regulatory abnormalities in which a process ends prematurely or too late.

Finally, a third kind of regulatory abnormality refers to abnormal rates with respect to a participant of the process that is being regulated. For example, a cytoplasmic flow rate can be increased or decreased not within a single *Cytoplasmic streaming* process but rather the total cytoplasmic flow rate, as a summation over all cytoplasmic streaming processes that occur within an organism (or a particular anatomical location), is increased or decreased. Although a flow rate of a single *Cytoplasmic streaming* process is a quality of that process, an increased *total* cytoplasmic flow rate is a quality of the regulation of *Cytoplasmic streaming*. In particular, it is possible for an organism to have a normal—or even a decreased—cytoplasmic flow rate in each individual cytoplasmic streaming process while at the same time having an increased total cytoplasmic flow rate due to a large increase in the frequency of occurrence of cytoplasmic streaming processes. Similarly, the frequency of occurrence of cytoplasmic streaming may be normal or decreased while the total cytoplasmic flow rate is increased due to an increased cytoplasmic flow rate in each individual cytoplasmic streaming process. We include total rates as subclasses of regulatory abnormalities in the CPO because these are the attributes of processes that are often measured or observed, while the rates of individual processes are inferred.

3.3 Implementation

We were faced with two choices for implementing the CPO: we could either implement a pre-composed ontology in which all classes and their definitions are pre-generated according to the patterns we define, or we could develop an annotation software that enables the selection of our process phenotype patterns based on the current structure of the GO. To maximize the utility and compatibility of the CPO, and to provide stable identifiers for its concepts, we selected the first strategy and developed a software to automatically generate a pre-composed ontology from the GO.

We developed a software that uses the OWL API (Horridge *et al.*, 2007) to generate an OWL representation of the CPO. The software requires three input files: a version of the GO on which to base the generated CPO, a version of PATO that is used to define abnormal qualities and a copy of the GO cross-product definitions (Mungall *et al.*, 2011) that is used to relate cell components to the processes in which they participate as well as identify the participants, inputs and outputs of processes.

We automatically generate a unique numerical identifier for each class in the CPO. Since the CPO is based on the GO and need to be updated with subsequent versions of the GO, we must ensure to keep identifiers stable in subsequent versions of CPO. Therefore, we use the identifiers for GO classes to generate CPO class identifiers.

In the CPO, identifiers contain two components and are of the form `CPO:XXGGGGGGG`, where `GGGGGGG` is the seven-digit identifier of the GO class on which the CPO class is based, and `XX` is a prefix that identifies the type of phenotype pattern that is applied to the GO class. For example, based on the class *Apoptosis* (`GO:0006915`), we generate the CPO classes *Abnormality of Apoptosis*, *Abnormality of single occurrence of apoptosis* and *Abnormality of regulation of apoptosis*. We use the prefixes `12`, 14 and 15 for each of the corresponding phenotype patterns and consequently generate the class identifiers CPO:120006915, CPO:140006915 and CPO:150006915. As long as the GO maintains its identifier for the *Apoptosis* class, the identifiers in the CPO will remain stable even when it is regenerated.

We use the labels of GO classes to automatically generate class labels for phenotype classes as well as textual definitions for classes in the CPO. For example, the label of the class for increased number of occurrences of *Apoptosis* is *Increased frequency of occurrences of Apoptosis*, and its textual definition states that an increased frequency of occurrences of *Apoptosis* is a phenotype of *Regulation of apoptosis* in which the number of occurrences of *Apoptosis* within a given time period is increased in comparison to a reference process that is considered *normal*.

As of November 2011, CPO contains 125 466 classes of which 79 236 are explicitly defined. The ELK reasoner (Kazakov *et al.*, 2011) is able to perform a classification of the ontology in under 10 s. We make the ontology and the source code that is used to generate it freely available on http://cell-phenotype.googlecode.com.

4 DISCUSSION

4.1 Applications of the CPO

The Fission Yeast Phenotype Ontology (FYPO), a new ontology developed to support annotation of phenotypes in *Schizosaccharomyces pombe*, consists of pre-composed terms describing normal or abnormal cellular phenotypes. Over 80% of FYPO definitions reference descendants of GO-BP's *Cellular process* as the entity; a further 11% reference GO-CC terms. All FYPO explicit definitions reference qualities in PATO, including *normal*, *abnormal* and several process qualities including *Increased duration* and *Decreased occurrence*. FYPO will, thus, fit neatly under the CPO umbrella and stands to benefit from the automated synchronization between CPO and GO, as well as the integration of cellular phenotypes across species that the CPO can provide. *S. pombe* annotations to FYPO terms will provide a rich body of highly specific, well-supported data to be integrated with data from other species.

A further domain that will greatly benefit from the CPO is *systems microscopy*, which aims to understand complex and dynamic cellular systems by combining automated fluorescence microscopy, cell microarray platforms, quantitative image analysis and data mining (Lock and Strmblad, 2010). If we consider some of the studies, which have been published in this field in the last few years (Fuchs *et al.*, 2010; Neumann *et al.*, 2010; Schmitz *et al.*, 2010), the need for CPO becomes evident. In the three studies, live-cell imaging assays and RNAi knockdown were used to generate phenotypic profiles that quantify the cellular response to a given siRNA thus allowing identification of hundreds of genes involved in diverse biological functions including cell division, migration and survival. In each study, several phenotypes were detected and described by the authors without the use of ontologies or controlled vocabularies, making the integration between datasets extremely difficult. For example, it is evident that cell division phenotypes were observed in all three datasets and referred to by terms such as *Mitotic delay/arrest*, *Prolonged mitotic exit*, *Methaphase delay* and *Methaphase cells*). Without a controlled vocabulary of cellular phenotypes, the overlap between such phenotype descriptions is unclear.

Data integration is also complicated by the lack of standardization at the level of data production and processing; all these issues are currently being address by the different groups involved in the Systems Microscopy Network of Excellence (http://www.systemsmicroscopy.eu/) and the first step towards data integration can be achieved by further developing CPO. This ontology will be used to integrate phenotypes' definitions across existing datasets and will then become an integrated part of the data processing pipeline and used to annotate the data as it gets generated (Conrad *et al.*, 2011).

4.2 Future research

Our main contribution is an analysis of process phenotypes that are used across multiple domains and scales and which are crucial for understanding and representing physiology of living systems. The Ontology of Physics in Biology (OPB) (Cook *et al.*, 2011) is an ontology that has recently been proposed to characterize physiological processes and the physical qualities of biological entities based on a theory of fluid dynamics. It is an important goal for future research to incorporate the OPB in phenotypic descriptions and make our theory of process phenotypes compatible with the physical descriptions of processes and their attributes as outlined by the OPB.

We implemented the CPO using a pattern-based approach to formulating phenotypes involving processes. The patterns we identify are based on pre-existing ontologies; in particular, the PATO ontology and the classification of cellular processes as well as CC in the GO. The result of our method is a large ontology in which classes for phenotypes are pre-composed: they are named and defined within an OWL ontology. However, the large size of the resulting ontology may impair its utility for data annotation and integration, and software tools may not always support such very large ontologies. The alternative to pre-composing all possible phenotype classes using the patterns we describe is to dynamically generate appropriately defined classes at the time at which they are being used. To achieve this goal, software must be developed to support ontology users in applying these patterns and generate the appropriate class description when required.

A further important task is to develop the theory we outlined and applied for the CPO. In particular, a precise formal characterization of this theory in terms of axioms will further improve the clarity of phenotypic descriptions of processes and enable its integration in well-developed formal ontologies of processes (Herre *et al.*, 2006; Özgövde and Grüninger, 2010).

ACKNOWLEDGEMENT

We thank Michel Dumontier for valuable discussions about the formal representation of the axioms and the different possible interpretations of the **regulates** relation.

Funding: Funding for R.H. is provided by the European Commission's 7th Framework Programme, RICORDO project, grant number 248502. M.A.H. is supported by Wellcome Trust, grant WT090548MA. H.H. is supported by the Institute for Medical Informatics, Statistics and Epidemiology, University of Leipzig. Funding for G.R. is provided by the European Union's Seventh

Framework Programme (FP7/2007-2013) under grant agreement number 258068, EU-FP7-Systems Microscopy NoE. G.V.G. is supported by the National Institute of Health, grant number R01 HG004838-02.

Conflict of Interest: none declared.

REFERENCES

Ashburner,M. *et al.* (2000) Gene ontology: tool for the unification of biology. *Nat. Genet.*, **25**, 25–29.

Bada,M. and Hunter,L. (2007) Enrichment of obo ontologies. *J. Biomed. Inform.*, **40**, 300–315.

Bada,M. *et al.* (2004) A short study on the success of the gene ontology. *Web Semantics: Science, Services and Agents on the World Wide Web*, **1**, 235–240.

Bard,J. *et al.* (2005) An ontology for cell types. *Genome Bio.*, **6**.

Berners-Lee,T. *et al.* (2001) The semantic web. *Scientific American*, **284**, 28–37.

Conrad,C. *et al.* (2011) Micropilot: automation of fluorescence microscopy-based imaging for systems biology. *Nat. Meth.*, **8**, 246–249.

Cook,D.L. *et al.* (2011) Physical properties of biological entities: An introduction to the ontology of physics for biology. *PLoS ONE*, **6**, e28708.

Eilbeck,K. *et al.* (2005) The sequence ontology: A tool for the unification of genome annotations. *Genome Biol.*, **6**.

Engel,S.R. *et al.* (2010) Saccharomyces genome database provides mutant phenotype data. *Nucleic Acids Res.*, **38**(Database issue): D433–D436.

Fuchs,F. *et al.* (2010) Clustering phenotype populations by genome-wide RNAi and multiparametric imaging. *Mol. Syst. Biol.*, **6**.

Gene Ontology Consortium (2010) The gene ontology in 2010: extensions and refinements. *Nucleic Acids Research*, **38**(Database issue), D331–D335.

Gkoutos,G.V. and Hoehndorf,R. (2011) Ontology-based cross-species integration and analysis of saccharomyces cerevisiae phenotypes. In *Proceedings of the 3rd Workshop for Ontologies in Biomedicine and Life sciences (OBML)*. Institute for Medical Informatics, Statistics and Epidemiology, Leipzig, Germany.

Gkoutos,G.V. *et al.* (2005) Using ontologies to describe mouse phenotypes. *Genome biology*, **6**.

Gkoutos,G.V. *et al.* (2009) Entity/quality-based logical definitions for the human skeletal phenome using PATO. *Annual International Conf. IEEE Eng. Med. Bio. Soc.*, **1**, 7069–7072.

Goble,C. and Stevens,R. (2008) State of the nation in data integration for bioinformatics. *J. Biomed. Inform.*, **41**, 687–693.

Grau,B. *et al.* (2008) OWL 2: The next step for OWL. *Web Seman.*, **6**, 309–322.

Gruber,T.R. (1995) Toward principles for the design of ontologies used for knowledge sharing. *Int. J. Hum. Compu. Stud.*, **43**.

Guarino,N. (1998) Formal ontology and information systems. In N. Guarino, editor, *Proceedings of the 1st International Conference on Formal Ontologies in Information Systems*, pp. 3–15. IOS Press.

Herre,H. *et al.* (2006) General Formal Ontology (GFO) – A foundational ontology integrating objects and processes [Version 1.0]. Onto-Med Report 8, IMISE, University of Leipzig, Leipzig, Germany.

Hoehndorf,R. *et al.* (2010a) Interoperability between phenotype and anatomy ontologies. *Bioinformatics*, **26**, 3112–3118.

Hoehndorf,R. *et al.* (2010b) Relations as patterns: Bridging the gap between OBO and OWL. *BMC Bioinformatics*, **11**, 441+.

Horridge,M. *et al.* (2007) Igniting the OWL 1.1 touch paper: The OWL API. In *Proceedings of OWLED 2007: Third International Workshop on OWL Experiences and Directions*.

Horrocks,I. (2007) OBO flat file format syntax and semantics and mapping to OWL Web Ontology Language. Technical report, University of Manchester.

Kazakov,Y. *et al.* (2011) Unchain my $\mathcal{EL}$ reasoner. In *Proceedings of the 23rd International Workshop on Description Logics (DL'10)*, CEUR Workshop Proceedings. CEUR-WS.org.

Lock,J.G. and Strmblad,S. (2010) Systems microscopy: an emerging strategy for the life sciences. *Exp. Cell Res.*, **316**, 1438–1444.

Motik,B. *et al.* (2009) Owl 2 web ontology language: Profiles. Recommendation, World Wide Web Consortium (W3C).

Mungall,C. *et al.* (2010a) Integrating phenotype ontologies across multiple species. *Genome Bio.*, **11**, R2+.

Mungall,C.J. *et al.* (2011) Cross-product extensions of the gene ontology. *J. Bio. Inform.*, **44**, 80–86.

Neumann,B. *et al.* (2010) Phenotypic profiling of the human genome by time-lapse microscopy reveals cell division genes. *Nature*, **464**, 721–727.

Ogren,P.V. *et al.* (2004) The compositional structure of gene ontology terms. *Pac. Symp. Biocomput.*, 214–225.

Özgövde,A. and Grüninger,M. (2010) Foundational process relations in bio-ontologies. In *Proceeding of the 2010 conference on Formal Ontology in Information Systems*, pages 243–256, Amsterdam, The Netherlands, The Netherlands. IOS Press.

Robinson,P.N. *et al.* (2008) The human phenotype ontology: a tool for annotating and analyzing human hereditary disease. *Am. J. Hum. Genet.*, **83**, 610–615.

Schindelman,G. *et al.* (2011) Worm phenotype ontology: integrating phenotype data within and beyond the c. elegans community. *BMC Bioinformatics*, **12**, 32.

Schmitz,M.H.A. *et al.* (2010) Live-cell imaging rnai screen identifies pp2ab55α and importin-β1 as key mitotic exit regulators in human cells. *Nature Cell Biology*, **12**, 886–893.

Schofield,P.N. *et al.* (2011) New approaches to the representation and analysis of phenotype knowledge in human diseases and their animal models. *Brief. Funct. Genomics*, **10**, 258–265.

Smith,B. *et al.* (2007) The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat. Biotech.*, **25**, 1251–1255.

Smith,C.L. *et al.* (2004) The mammalian phenotype ontology as a tool for annotating, analyzing and comparing phenotypic information. *Genome Bio.*, **6**, R7.